AI-based Translator for Languages with Limited Resources

Jordan Alexander[1], Saad Hasan[1], Omair Shafiq[1], Ali Arya[1], Teresa Samson[2]

[1] Carleton University, Ottawa, ON K1S5B6, Canada
[2] The First Nation of Na-Cho Nyäk Dun, Mayo, YT Y0B 1M0, Canada

Corresponding author: Ali Arya, arya@carleton.ca

Abstract. Neural Machine Translation (NMT) is being explored for low-resource language support, due to its potential to assist language learning and revitalization. However, developing translation systems for Indigenous languages remains challenging due to limited data, poor tokenizer coverage, language complexity, and data stewardship. To address these limitations, this work proposes a parameter-efficient pipeline for Northern Tutchone and English using ByT5-small with Low-Rank Adaptation (LoRA). ByT5-small uses byte-level sequence-to-sequence modeling to avoid a language-specific tokenizer and represent Northern Tutchone orthographic features. LoRA trains low-rank adapter weights while keeping most pretrained parameters frozen. The study compares 1.5k and 4.9k sentence-pair datasets across two directions and three evaluation seeds. Performance was evaluated using BLEU, chrF++, and BLEURT, with paired bootstrap resampling and paired t-test used to assess significance. Results demonstrate that dataset expansion improved Northern Tutchone to English translation, increasing mean BLEU from 14.95 to 26.93, chrF++ from 25.03 to 37.82, and BLEURT from 0.2936 to 0.4057. English to Northern Tutchone remained more difficult, with BLEU increasing from 4.16 to 5.19 and chrF++ from 13.27 to 21.37. The primary contribution is the empirical investigation of a ByT5-LoRA translation approach for this language pair, establishing baseline results in a low-resource Indigenous language setting.

Keywords: Northern Tutchone, Low-Resource Machine Translation, Indigenous Language Technology, ByT5, LoRA, Language Revitalization.

1 Introduction

Neural Machine Translation (NMT) has become increasingly capable for high-resource language pairs, due to large parallel corpora, pretrained sequence-to-sequence models, and parameter-efficient fine-tuning strategies that allow pretrained models to adapt to specific languages and translation tasks [1-3]. These systems show great promise for supporting language learning, documentation, and interactive educational tools, with the ultimate goal of language and cultural preservation, particularly when these tools can be integrated into broader human-centered workflows. However, there are

significant challenges in developing translation systems for Indigenous and endangered languages. Many of these languages have limited digitized text, very small parallel corpora, limited representation in pretrained language models, and important requirements around language data stewardship and data governance [4-7]. Furthermore, standard NMT systems often rely on subword tokenizers trained on high-resource languages, which can poorly represent low-resource languages. This creates a major limitation for languages such as Northern Tutchone, where the available data is scarce and where accurate representation of character-level structure is critical.

Northern Tutchone is the language spoken by the First Nation of Na-Cho Nyäk Dun (FN-NND), Yukon, Canada, with limited available parallel text for machine learning. This data scarcity makes conventional high-data NMT approaches difficult to apply directly, because model performance depends heavily on the size, quality, and representativeness of the available corpus. An additional challenge arises from the translation direction itself. Translating from Northern Tutchone to English requires the model to interpret a low-resource source language and generate English, a high-resource target language that is well represented in pretrained models. However, translating from English to Northern Tutchone is more difficult because the model must generate the low-resource language directly, including its orthographic patterns, diacritics, and word forms, from a limited number of examples. This direction asymmetry creates an important evaluation problem, as improvements may appear differently across metrics and may not be captured equally by word-level and character-level measures.

To address the data and compute limitations of this setting, this work used a ByT5-small translation pipeline adapted with Low-Rank Adaptation (LoRA). ByT5 is a byte-level encoder-decoder transformer, meaning that text is represented directly at the byte level rather than through a fixed subword vocabulary [8]. This was useful for Indigenous languages such as Northern Tutchone because the model did not require a custom tokenizer and could directly represent non-ASCII characters and diacritics. LoRA was used to reduce the number of trainable parameters by inserting trainable low-rank adapters into selected attention layers while leaving the majority of the pretrained model frozen [3]. This made the approach more practical for a small research workflow with limited computational resources.

The central question examined in this paper was how much translation performance improved when the available parallel training data was increased from approximately 1.5k sentence pairs to approximately 4.9k sentence pairs. This question is important because data collection and preparation are often among the most difficult parts of low-resource Indigenous-language technology work. If a relatively modest expansion in cleaned parallel data produces large performance gains, this supports continued investment in data collection and preparation. However, if gains are uneven across translation directions, this also provides important design guidance for how the system should and should not be used in Human-Computer Interaction (HCI) settings.

This study compares a smaller parallel dataset with a larger, expanded dataset across two translation directions and three random seeds. Performance is evaluated using BLEU, chrF++, and BLEURT, with paired bootstrap resampling used to test whether improvements from the larger dataset are statistically meaningful [9-12]. This experimental design allows the effect of dataset expansion to be evaluated while accounting

for seed-level variability and direction-specific differences in translation difficulty. Because the two translation directions are not equivalent tasks, they were evaluated separately rather than as a single bidirectional result.

The primary contribution of this work is the demonstration of a feasible and statistically evaluated parameter-efficient translation workflow for a very-low-resource Indigenous language pair, providing the first baseline results. A secondary contribution is the interpretation of these results for HCI design. Specifically, the results suggest that the current model should be treated as a support component within constrained, human-in-the-loop interactions, rather than as an autonomous translator. The paper therefore connects model development and evaluation with practical design implications for future language learning systems. Our work is part of a language revitalization project by the FN-NND that uses artificial intelligence, holographic displays, and VR to create a multi-platform educational record of an Indigenous culture (https://kwandekando.ca). The project is in partnership with Carleton University (Ottawa, Canada) and the Office of the Commissioner of Indigenous Languages.

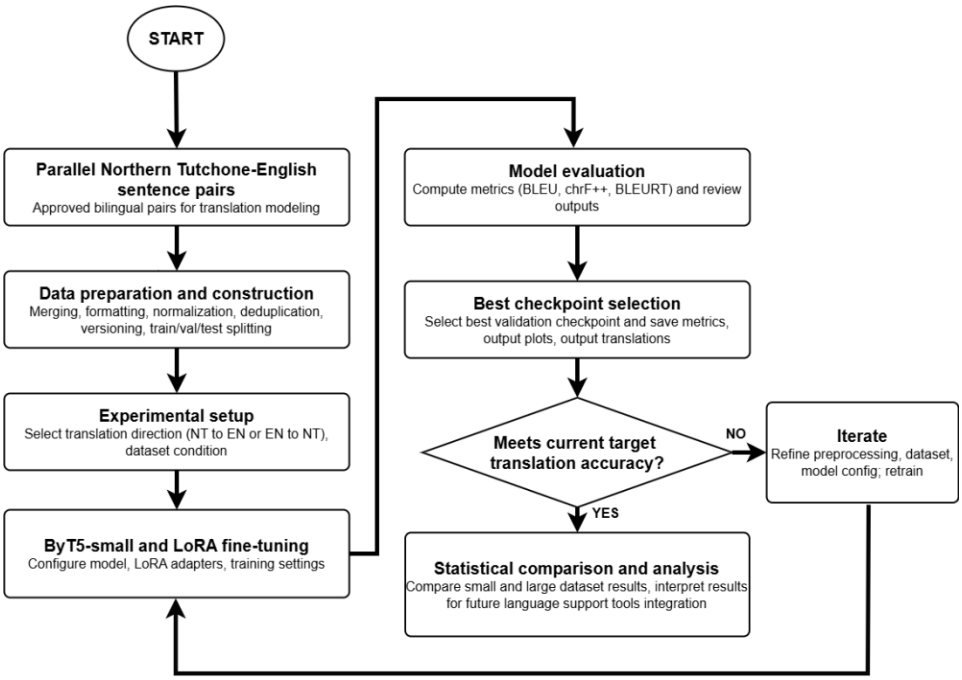

Fig. 1. Overview of the Northern Tutchone and English translation development pipeline.

2 Related works

Transformer-based sequence-to-sequence models provide the foundation for modern NMT by enabling attention-based modeling of long-range relationships in text [1]. Building on this architecture, text-to-text transfer learning frameworks such as T5 showed that pretrained encoder-decoder models can be adapted across a wide range of language tasks [2]. However, conventional fine-tuning can require substantial data and computational resources, which is a major limitation in low-resource settings. Parameter-efficient methods such as LoRA address this by training low-rank adapter weights while keeping most pretrained parameters frozen, reducing the cost of adaptation while still allowing the model to specialize to a new task [3]. For low-resource languages, byte-level modeling provides another important advantage by representing text directly at the byte level rather than relying on a fixed subword vocabulary [8]. This is particularly relevant for languages with limited tokenizer coverage, orthographic variation, and non-ASCII characters.

Previous work on Indigenous language technologies has emphasized that technical development must account for limited data availability, language-specific representation issues, and the broader social context of language data use [4-7]. Mager et al. describe the challenges of building language technologies for Indigenous languages of the Americas, including limited resources and the need for approaches that fit low-data conditions [4]. Bird and Cooper et al. further emphasize that language technology work should attend to process, community context, and responsible development rather than treating language data as a generic technical resource [5, 7]. In the Canadian First Nations context, OCAP principles also highlight the importance of ownership, control, access, and possession in data governance [6]. Together, this previous work supports the need for translation approaches that are data-efficient, tokenizer-robust, and developed with attention to language stewardship, rather than treating Indigenous language translation as a standard high-resource NMT problem.

3 Methods

This work implemented a low-resource NMT workflow that combines a byte-level sequence-to-sequence model, parameter efficient fine-tuning, applied to a controlled dataset comparison with statistical testing. The workflow was designed to evaluate the translation accuracy of a low-resource Indigenous language and test whether increasing the size of the cleaned parallel corpus improved translation performance while maintaining a consistent training and evaluation procedure. The experimental design compared two dataset sizes, two translation directions, and three random seeds, producing twelve training runs in total.

The datasets consisted of parallel Northern Tutchone and English sentence pairs. The smaller dataset contained approximately 1.5k parallel sentence pairs and represented the earlier dataset used in the project. The larger dataset contained approximately 4.9k sentence pairs and represented an expanded dataset curated and prepared later in the project. The larger dataset was approximately 3.3 times the size of the smaller

dataset. Both datasets were evaluated in Northern Tutchone to English and English to Northern Tutchone directions separately. These directions were treated as separate tasks because they placed different demands on the model. Northern Tutchone to English required the model to interpret Northern Tutchone input and generate English output. English to Northern Tutchone required the model to generate Northern Tutchone output, which was expected to be more difficult because the target language had much less representation in the pretrained model. Three random seeds were used for each condition: 0, 43, and 99. This provided a more stable performance estimate than a single run and allowed the results to show whether improvements were consistent across different splits and initialization conditions.

Since the experiments were conducted in a very-low-resource language setting, data preparation was central to the reliability of the evaluation. The data was cleaned and normalized to reduce formatting inconsistencies, remove duplicate entries, and improve consistency between parallel pairs. This process was especially important because small datasets are more vulnerable to inflated or unstable evaluation results when repeated or near-duplicate examples appear across training and validation sets.

The data preparation process also accounted for the structure and distribution of the available language data. The curated low-resource corpora often contained short 1-2 word phrases, and while these examples can be useful for training, they can distort evaluation if they dominate the validation set, as it makes it more unclear if the model is learning sentence structure or remembering single words/phrases. For this reason, sentence pairs with source sentences of two words or fewer were preferentially routed to the training set. This kept more of the very short and potentially less informative examples out of evaluation, while still allowing the model to learn from them during training, and instead filling the evaluation sets with more 3+ word phrases/sentences.

Each seed of the small dataset was split into train, validation, and test subsets using an 80/10/10 ratio. The splitting procedure was cluster-aware, where semantically similar sentence pairs were grouped together and assigned to the same split, which reduced the risk of near-duplicate pairs appearing on both sides of the train and validation data splits, and this prevents data leakage. The small and large dataset models were evaluated on the same validation references generated by the seed splits of the smaller dataset, containing approximately 184 sentences. This shared-reference structure was critical because it allowed the statistical analysis to compare model outputs sentence-by-sentence rather than treating each system as an unrelated aggregate score. Because the data is connected to community language resources, the paper does not reproduce the full dataset or provide unrestricted access to sentence-level content. Instead, the methodology reports the dataset sizes, split logic, training setup, and evaluation procedure, to provide technical transparency and respect language data stewardship requirements.

All experiments used Google's ByT5-small as the base model [8]. ByT5-small is a byte-level encoder-decoder transformer with approximately 300 million parameters. The byte-level architecture was selected because it does not require a language-specific tokenizer. This was important for Northern Tutchone, where the small corpus size, diacritics, and possible spelling variation made it difficult to rely on a conventional subword tokenizer with no Northern Tutchone representation. By representing text at the

byte level, the model could process Northern Tutchone characters directly without out-of-vocabulary tokenization issues.

The base model was adapted using LoRA [3]. LoRA inserts trainable low-rank matrices into selected layers of the pretrained model while keeping the majority of model parameters frozen. This reduces the number of trainable parameters and makes the training process more efficient. In this work, LoRA was conservatively applied to the query and value projection layers in the attention mechanism, corresponding to the q and v modules in ByT5's encoder-decoder architecture. The LoRA rank was set to 64, lora_alpha was set to 128, and dropout was set to 0.1. This configuration provided a relatively high-capacity adapter while maintaining the parameter-efficient structure of the fine-tuning process.

Table 1. Model and training configurations.

Component	Configuration
Base model	ByT5-small
Adaptation method	LoRA
LoRA rank	64
LoRA alpha	128
LoRA dropout	0.1
Target modules	Query and value projection layers
Optimizer	paged AdamW 8-bit
Learning rate	0.001
Scheduler	Linear decay
Label smoothing	0.1
Maximum epochs	200
Beam size (generation)	4
Checkpoint reported	Best validation checkpoint
Seeds	0, 43, 99
Datasets	Small (~1.5k samples), Large (~4.9k samples)

The training configuration was held consistent across the dataset-size comparison. Models were trained for up to 200 epochs without patience, allowing the training process to complete a full sweep rather than stopping early after a small number of stagnant validation epochs, allowing for a clear comparison between results. A learning rate of 0.001 was used with a linear decay scheduler, and label smoothing was set to 0.1. The optimizer was paged AdamW 8-bit, which reduced memory requirements during optimization. Additional memory optimizations included mixed precision training and gradient checkpointing. Validation generation used beam search with 4 beams. The best validation checkpoint from each run was used for final reporting, meaning that the reported scores represent the strongest validation checkpoint observed during training rather than the final epoch. The validation set was used for checkpoint selection and model comparison. The test set was held out from training and checkpoint selection,

and was used only as a post-hoc check that the selected models showed similar behavior on the unseen data.

Model performance was evaluated using BLEU, chrF++, and BLEURT. BLEU was included as a standard corpus-level machine translation metric based on n-gram/word-level overlap [9]. chrF++ was included because it evaluates character-level similarity while also incorporating word n-gram information [10]. This makes it especially useful in a low-resource and morphologically complex setting, where a model may begin learning meaningful surface-form patterns even when exact word-level overlap is limited. BLEURT was included for the Northern Tutchone to English direction as a learned reference-based metric that provided an additional signal of English output quality [11].

BLEU and chrF++ were computed using SacreBLEU [13]. BLEURT was computed using BLEURT-20 through the HuggingFace evaluate package. Unicode NFC normalization was applied to all predictions and references before scoring to ensure consistency across both metrics and the bootstrap resampling procedure. BLEURT was only reported for Northern Tutchone to English generation, because the metric requires a model for the target language, and no BLEURT model exists for Northern Tutchone. Therefore, BLEURT could not be computed for English to Northern Tutchone generation.

Paired bootstrap resampling was used to test whether the larger dataset improved BLEU and chrF++ compared to the smaller dataset [12]. For each seed, translation direction, and metric, the small and large models were evaluated over the same validation references. Sentence-level outputs were resampled with replacement 10,000 times, and the metric difference between the large and small dataset models was recomputed for each resample using the same SacreBLEU calls applied during training, ensuring no deviation in metric computation between the training evaluation and the bootstrap procedure. The minimum reportable p-value was bounded at 1/10,000 due to the number of resamples. This preserved the pairing between systems and estimated how consistently the observed improvement appeared under resampling of the validation set. Because the three seeds used different splits, the per-seed bootstrap p-values were combined using Fisher's method rather than pooling all outputs as if they came from a single identical evaluation set [14]. BLEURT was evaluated differently because the available analysis used aggregate seed-level BLEURT scores rather than live per-sentence BLEURT resampling. For this reason, BLEURT was compared using a two-tailed paired t-test across the three seed-level scores in the Northern Tutchone to English direction.

4 Results

Increasing the dataset size from approximately 1.5k to 4.9k sentence pairs improved translation performance in both directions, but the strength and consistency of the improvement depended on the translation direction and evaluation metric. The largest and most consistent gains occurred for Northern Tutchone to English, where all three metrics improved across all three seeds. English to Northern Tutchone remained more

difficult, with low BLEU scores and greater variability, but chrF++ improved substantially and significantly.

For Northern Tutchone to English translation, the larger dataset improved performance across every seed and every metric. Mean BLEU increased from 14.95 with the small dataset to 26.93 with the large dataset, corresponding to an 80.1% relative improvement. Mean chrF++ increased from 25.03 to 37.82, corresponding to a 51.1% relative improvement. Mean BLEURT increased from 0.2936 to 0.4057, corresponding to a 38.2% relative improvement. The BLEU and chrF++ improvements were highly significant under paired bootstrap resampling, with combined p-values below 0.0001. This indicated that the larger dataset produced improvements that were not only visible in the aggregate means, but also stable under sentence-level resampling of the validation set. BLEURT also improved for all three seeds, but the paired t-test over seed-level aggregates gave $p = 0.0501$, narrowly missing the conventional $p < 0.05$ threshold.

For English to Northern Tutchone translation, the larger dataset also improved the mean scores, but the results were less consistent. Mean BLEU increased from 4.16 to 5.19, corresponding to a 24.6% relative improvement, but this difference was not statistically significant under paired bootstrap resampling. The p-value for English to Northern Tutchone BLEU was 0.0950. The per-seed BLEU values were also less stable than in the Northern Tutchone to English direction. In contrast, mean chrF++ increased from 13.27 to 21.37, corresponding to a 61.0% relative improvement. This improvement was highly significant, with a combined p-value below 0.0001. These results indicate that the larger dataset improved character-level similarity to the Northern Tutchone references, even where exact word-level overlap remained limited. In practice, this suggests the model produced translations which were closer to correct but which contained typo-level errors. The main quantitative results and significance tests are summarized in Table 2.

Table 2. Main quantitative results and significance

Direction	Metric	Mean (1.5k)	Mean (4.9k)	% Increase	p-value
NT to EN	BLEU	14.95	26.93	+80.1%	<0.0001
NT to EN	chrF++	25.03	37.82	+51.1%	<0.0001
NT to EN	BLEURT	0.2936	0.4057	+38.2%	0.0501
EN to NT	BLEU	4.16	5.19	+24.6%	0.0950
EN to NT	chrF++	13.27	21.37	+61.0%	<0.0001

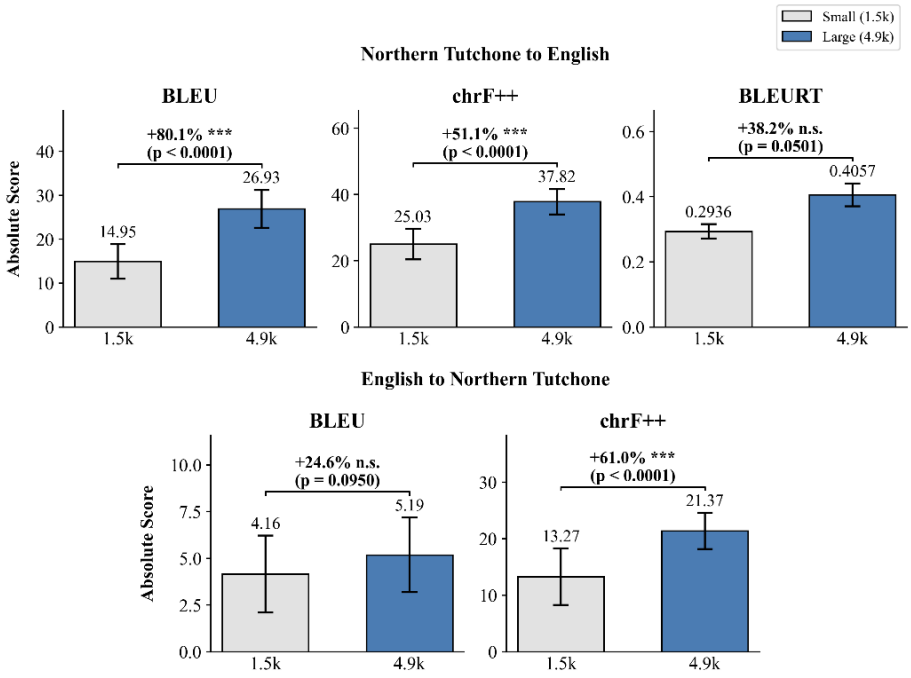

Figure 2. Effect of dataset size on translation performance across direction and metric. Significance markers indicate the statistical significance of the improvement from the small to large dataset, where *** indicates $p < 0.0001$ and n.s. indicates not significant.

Figure 2 shows the same results visually, with mean validation scores from the best validation checkpoint for each run, across three seeds for the small (1.5k) and large (4.9k) datasets. Error bars represent ±1 standard deviation across seeds. BLEU and chrF++ p-values were computed using paired bootstrap resampling with Fisher-combined p-values, while BLEURT was evaluated using a paired t-test over seed-level aggregate scores.

The difference between BLEU and chrF++ was especially important for interpreting the English to Northern Tutchone results. BLEU depends on exact word-level n-gram overlap, which can be difficult to achieve in low-resource generation with limited references. chrF++ is more sensitive to character-level overlap and can capture partial improvements in spelling patterns, stems, endings, and orthographic structure. Therefore, the English to Northern Tutchone results should not be interpreted as showing no improvement. Instead, they suggest that the larger dataset helped the model learn more Northern Tutchone surface structure, but that this improvement was not yet strong enough to produce consistent fully accurate translations word-for-word.

The best absolute metric scores were also achieved by the large dataset condition. For Northern Tutchone to English, the highest BLEU score was 31.82, the highest chrF++ score was 42.15, and the highest BLEURT score was 0.4287, all from the large dataset condition with seed 0. For English to Northern Tutchone, the highest BLEU

score was 7.48 and the highest chrF++ score was 24.69, also from the large dataset condition with seed 0.

Table 3: Summary of best absolute metric scores for each metric, from all investigated training experiments.

Direction	Metric	Best	Condition
NT to EN	BLEU	31.82	Large dataset, Seed 0
NT to EN	chrF++	42.15	Large dataset, Seed 0
NT to EN	BLEURT	0.4287	Large dataset, Seed 0
EN to NT	BLEU	7.48	Large dataset, Seed 0
EN to NT	chrF++	24.69	Large dataset, Seed 0

Representative validation examples further illustrate how the larger dataset changed model behavior. Table 4 compares outputs from the small and large dataset models using validation examples from seed 0. These examples were selected to show specific cases of clear improvement, partial improvement, and errors observed in the validation sets evaluated by the model at the best checkpoints.

For Northern Tutchone to English, the larger dataset produced examples where the small model generated an unrelated English sentence, while the large model recovered the full translation meaning. For example, the Northern Tutchone input “Dek'ą łądézhe.” was translated by the small dataset model as “The moose is sitting,” while the large dataset model accurately produced “The man is going hunting,” matching the reference. The second example showed a partial improvement: the small dataset model translated “Kensun t'éhe'yó.” as “The boy is sitting,” while the large dataset model produced “He is wearing warm moccasins.” This recovered the main meaning, but introduced an additional detail not present in the reference. Example 2 also shows a limitation in the reference-based evaluation, as the large-dataset model produced the correct English spelling of “moccasins” while the reference contained the misspelling “moccassins”, reducing the model’s BLEU and chrF++ scores. This highlights the fact that automatic quantitative metrics can be negatively affected by spelling errors in the reference dataset.

For English to Northern Tutchone, the examples show both the promise and limitations of the larger dataset. In example 3, the input “I am making a fire” was translated incorrectly by the small dataset model, while the large dataset model fully matched the Northern Tutchone reference. In example 4, the input “I want meat” produced a mixed improvement. The small dataset output captured an essentially correct second word/phrase with a typo-level error, but used an incorrect first word/phrase, while the large dataset output captured the first word/phrase but it generated an incorrect second word/phrase. This example shows a mixed improvement pattern, where the larger dataset improved one part of the translation while introducing an error in the other.

Table 4. Representative validation output comparison from Seed 0.

Example	Output comparison
1. NT to EN	**Input:** Dek'ą́ łądézhe. **Reference:** The man is going hunting. **Small dataset:** The moose is sitting. **Large dataset:** The man is going hunting.
2. NT to EN	**Input:** Kensun t'éhe'yó. **Reference:** He is wearing moccassins [sic]. **Small dataset:** The boy is sitting. **Large dataset:** He is wearing warm moccasins.
3. EN to NT	**Input:** I am making a fire. **Reference:** Kwǻ́n díhk'ę́. **Small dataset:** Edu iht'in. **Large dataset:** Kwǻ́n díhk'ę́.
4. EN to NT	**Input:** I want meat. **Reference:** Etthǻ́n yenihthän. **Small dataset:** Ts'ǻ́t yenithän. **Large dataset:** Etthǻ́n seétlę.

5 Discussion

The results show that increasing the amount of cleaned parallel training data had a substantial positive effect on Northern Tutchone and English translation performance. The clearest improvements occurred for the Northern Tutchone to English direction, where the larger dataset produced strong and statistically significant gains in both BLEU and chrF++. The improvement from 14.95 to 26.93 mean BLEU represents a large practical gain in this low-resource setting, while the mean chrF++ improvement from 25.03 to 37.82 confirms that the outputs became more similar to the references at both word and character levels. These results support the value of continued corpus development and data preparation, especially when paired with consistent evaluation splits and statistical testing.

However, the results also show that dataset expansion at this small scale alone does not make the system ready for unrestricted use. English to Northern Tutchone remained substantially more difficult than Northern Tutchone to English. The BLEU improvement in this direction was modest and not statistically significant, even though chrF++ improved significantly. This suggests that the larger dataset helped the model learn more Northern Tutchone surface structure, but did not yet enable consistently accurate full-sentence translations. This distinction is important because English to Northern Tutchone is the direction most likely to be used for learner-facing translations, phrase

practice, or draft language production. Errors in this direction may be more harmful if users interpret model outputs as authoritative. Conversely, Northern Tutchone to English translation may be more appropriate for comprehension support, review workflows, or assisting users in navigating existing Northern Tutchone text.

The difference between BLEU and chrF++ also provides an important interpretation of model progress. BLEU is useful as a standard machine translation metric, but it can be harsh in low-resource settings because it rewards exact n-gram overlap [9]. This is especially limiting when there are few references and when multiple valid translations or spelling variants may exist. chrF++ provides a complementary signal by measuring character-level similarity with additional word n-gram information [10]. In this study, chrF++ captured a strong improvement for English to Northern Tutchone even when BLEU did not. This suggests that the model began learning more of the target-language form, but that word-level accuracy and full translation correctness remained limited.

At the same time, chrF++ should not be overinterpreted. A higher character-level score does not guarantee semantic correctness, grammatical acceptability, or cultural appropriateness. A model may produce an output that is closer to the reference at the character level while still being wrong in meaning – for example, outputting “mouse” instead of “moose”. This is especially important for HCI applications, where the practical concern is not only whether a metric improves, but whether a learner, teacher, or reviewer can safely use the output. The representative examples in Table 4 show this issue directly: the larger dataset often produced outputs closer to the reference, but some outputs still contained word-level errors that would require human review.

The results also support the practical value of using ByT5-small with LoRA in this setting. ByT5’s byte-level representation was well suited to Northern Tutchone because it avoided training a tokenizer and allowed the model to process orthographic features and diacritics directly [8]. LoRA reduced the training burden by limiting the number of trainable parameters and allowing adaptation through query and value projection adapters [3]. This is important for low-resource and community-partnered language technology projects, where access to large compute infrastructure may be limited. In practice, the larger dataset runs required approximately 3.7 hours per run, while the smaller dataset runs required approximately 1.0 to 1.2 hours, both on a single NVIDIA GeForce RTX 3070 Ti graphics card, indicating that the expanded dataset produced meaningful gains while remaining feasible for repeated experimentation.

From an HCI perspective, the current system is best understood as a developing translation support component that can be integrated into human-in-the-loop workflows. The results are strong enough to support continued development toward practical language support tools, particularly where model outputs can be reviewed, corrected, or constrained to appropriate learning contexts. A correction-based interface would allow the model to generate a draft translation that a learner, reviewer, or language worker could accept, reject, or correct. These corrections could then be stored for future review and potential retraining, creating a feedback loop between model use, language review, and dataset improvement. This positions the model as a practical foundation for future community-driven language learning systems, where translation assistance can be provided while still maintaining human oversight and data stewardship.

A second appropriate use case is constrained translation within language learning environments. Instead of allowing unrestricted free-text translation, the system could be limited to reviewed prompts, known vocabulary sets, or scripted learning scenarios. This would reduce the risk of out-of-distribution outputs and allow the system to support interaction without presenting unverified translations as authoritative. For example, a web-based or immersive learning activity could use the model to support controlled phrase practice, but only within content areas that have been reviewed or constrained by language experts.

The project also raises important governance and responsible-use considerations. In many machine learning settings, reproducibility is framed mainly as releasing code, datasets, and model weights. In Indigenous language technology work, reproducibility must be balanced with language data stewardship and community governance. The datasets can contain culturally sensitive and personally identifiable information and is connected to language resources that require appropriate control and review. For this reason, this paper reports the model configuration, training setup, evaluation process, and statistical methods, while avoiding unrestricted disclosure of the underlying language data.

There are several limitations to this study. First, due to the inherent data scarcity of this project, the dataset remained very small compared with typical neural machine translation corpora. Although the larger dataset contained approximately 4.9k sentence pairs, this is still extremely limited for training a translation system, especially for generation into Northern Tutchone. Second, the evaluation relied primarily on automatic metrics. BLEU, chrF++, and BLEURT provide useful quantitative signals, but they do not fully measure semantic correctness, grammatical acceptability, cultural appropriateness, or learner safety. The translation outputs were not evaluated through a comprehensive human linguistic review, partly due to the number of outputs for review, and the need for Northern Tutchone language fluency and expertise to accurately assess correctness. Third, due to computational limitations, the study used three random seeds. This is stronger than reporting a single run, but it is still limited for estimating the full variability of training. The BLEURT comparison was especially limited because it relies on an evaluation model trained on the target language, of which none exists for Northern Tutchone, thus could only be evaluated for one direction. Finally, the study did not include a user evaluation. Future HCI work will be needed to determine how learners, teachers, and reviewers interact with model outputs in realistic interface settings.

Overall, these results indicate that the larger dataset substantially improved model performance, but that the model at its current capability should remain embedded within human review and constrained interaction design. Until further improvements can be gained through deeper expansion of the dataset, the translation system is therefore currently most useful as a foundation for future language support workflows, where model outputs can assist learners and reviewers without replacing human expertise.

6 Conclusion

Overall, this work demonstrated a feasible, parameter-efficient workflow for Northern Tutchone and English neural machine translation using ByT5-small with LoRA. Starting from a very small parallel corpus, the study evaluated whether expanding the dataset from approximately 1.5k to 4.9k sentence pairs improved translation performance across two directions and three random evaluation seeds. Results showed that increasing the amount of parallel data produced substantial improvements, particularly for Northern Tutchone to English translation, where mean BLEU improved from 14.95 to 26.93 and mean chrF++ improved from 25.03 to 37.82, with both improvements statistically significant. English to Northern Tutchone remained more difficult, with BLEU improving only modestly, while chrF++ improved significantly from 13.27 to 21.37. These results indicate that dataset expansion improved both translation directions, but that Northern Tutchone generation remained more challenging than English generation.

These findings matter because they show that a practical translation workflow can be developed even in a very-low-resource Indigenous language setting, while also identifying the conditions where human review remains important. From a HCI perspective, the system provides a foundation for future language support tools where translation outputs can be reviewed, corrected, and improved through human-in-the-loop workflows. Future work should expand the parallel dataset, evaluate additional model architectures and adaptation strategies, incorporate structured human evaluation, and develop interface prototypes that support uncertainty, review, and correction while respecting language data stewardship.

Acknowledgments. This work is part of a language revitalization project by the First Nation of Na-Cho Nyäk Dun, co-developed by Carleton University. The authors wish to thank the community members and Elders in Mayo, Yukon, whose participation, knowledge, and generosity made this research possible. We are grateful to all who shared their stories and time with us. The project is funded by the Office of the Commissioner of Indigenous Languages.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30, pp. 5998-6008. Curran Associates, Inc. (2017).
2. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21(140), 1-67 (2020).
3. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022).

4. Mager, M., Gutierrez-Vasques, X., Sierra, G., Meza-Ruiz, I.: Challenges of language technologies for the Indigenous languages of the Americas. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 55-69. Association for Computational Linguistics, Santa Fe, New Mexico (2018).
5. Bird, S.: Decolonising speech and language technology. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 3504-3519. International Committee on Computational Linguistics, Barcelona (Online) (2020).
6. First Nations Information Governance Centre: The First Nations Principles of OCAP®. https://fnigc.ca/ocap-training/, last accessed 2026/05/27.
7. Cooper, N., Heldreth, C., Hutchinson, B.: "It's how you do things that matters": Attending to process to better serve Indigenous communities with language technologies. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 204-211. Association for Computational Linguistics, St. Julian's, Malta (2024).
8. Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., Raffel, C.: ByT5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the Association for Computational Linguistics 10, 291-306 (2022).
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318. Association for Computational Linguistics, Philadelphia (2002).
10. Popović, M.: chrF++: Words helping character n-grams. In: Proceedings of the Second Conference on Machine Translation, pp. 612-618. Association for Computational Linguistics, Copenhagen (2017).
11. Sellam, T., Das, D., Parikh, A.: BLEURT: Learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7881-7892. Association for Computational Linguistics, Online (2020).
12. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Lin, D., Wu, D. (eds.) Proceedings of EMNLP 2004, pp. 388-395. Association for Computational Linguistics, Barcelona (2004).
13. Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 186-191. Association for Computational Linguistics, Brussels (2018).
14. Fisher, R.A.: Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh (1925).